

COGNITIVE-AWAKENING CHAIN-OF-SURGERY FOR COMPOSITIONAL ZERO-SHOT SURGICAL TRIPLET RECOGNITION (SUPPLEMENTARY)

Anonymous authors

Paper under double-blind review

1 TRIPLET CATEGORY SPLITTING

We use $\gamma = 80$ in the paper, which is a suitable setting for both training and evaluation on CholecT50. As shown in Table 4, a total of 60 triplet types are used for training, with 40 compositional-unseen categories held out for evaluating and testing.

2 CHAIN-OF-SURGERY PROMPTING TEXT

The prompts used in CoS are shown in Figure 4. In the gaze stage, thinking contents are concatenated with the answers generated from the glance stage directly as the assistant prompt. In the think stage, answers generated from the gaze stage are used as assistant prompts separately.

3 IMPLEMENTATION DETAILS

We adopt the AdamW optimizer with a learning rate of 0.001 and a weight decay of 0.0001. Binary Cross-Entropy loss is used for instrument, verb, target, and triplet recognition, while an additional InfoNCE loss is applied between the text representations produced at the Gaze stage and the Video Encoder features. The batch size is set to 4, corresponding to four video clips per batch. Both the Frame Encoder and Video Encoder are trained with an early-stopping strategy, using a patience of 5 and a minimum delta of 0.005.

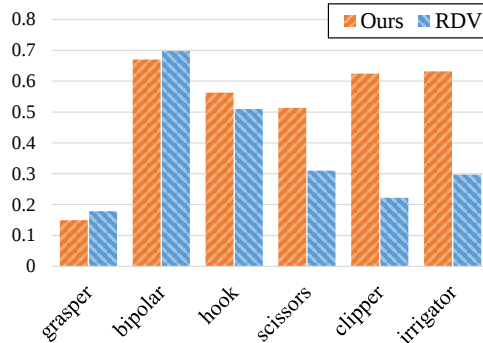


Figure 1: Instrument improvements in mAP_i of our framework compared to the RDV (the second-best method).

Method	mAP _i	mAP _v	mAP _t	mAP _{ivt}
Glance	13.62	15.54	9.24	2.11
Glance+Gaze	19.07	17.85	9.47	2.35
Glance+Gaze+Think	19.88	19.26	10.52	2.51
Glance+Gaze+Think+Machine Encoding	32.82	27.67	22.30	9.98

Table 1: Ablation study for Chain-of-Surgery (CoS) with Qwen3-VL-8B-Instruct&Qwen3-VL-8B-Thinking setting.

clip length T	mAP _i	mAP _v	mAP _t	mAP _{ivt}
8	49.70	32.36	20.15	3.31
16	52.54	37.56	21.55	9.39
32	51.41	37.14	20.72	9.26

Table 2: Ablation study for clip length T

REFERENCES

- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Yiliang Chen, Zhixi Li, Cheng Xu, Alex Qinyang Liu, Xuemiao Xu, Jeremy Yuen-Chun Teoh, Shengfeng He, and Jing Qin. Prostatd: A large-scale multi-source dataset for structured surgical triplet detection. *arXiv preprint arXiv:2506.01130*, 2025.
- Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022.
- Saurav Sharma, Chinedu Innocent Nwoye, Didier Mutter, and Nicolas Padoy. Rendezvous in time: an attention-based temporal fusion approach for surgical triplet recognition. *International Journal of Computer Assisted Radiology and Surgery*, 18:1053–1059, 2023.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Hecvl: hierarchical video-language pretraining for zero-shot surgical phase recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 306–316, 2024a.
- Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. *Advances in Neural Information Processing Systems*, 37:122952–122983, 2024b.
- Kun Yuan, Vinkle Srivastav, Tong Yu, Joel L Lavanchy, Jacques Marescaux, Pietro Mascagni, Nassir Navab, and Nicolas Padoy. Learning multi-modal representations by watching hundreds of surgical video lectures. *Medical Image Analysis*, pp. 103644, 2025.

Algorithm 1 SAM-based Region Selection Pipeline

```

1: Input: Image  $I$ , top- $k$ 
2: Output: Selected masks
3:  $\mathcal{S} \leftarrow \text{SELECTREGIONS}(I, k)$ 
4: return  $\mathcal{S}$ 

5: function SELECTREGIONS( $I, k$ )
6:    $\mathcal{M} \leftarrow \text{SamAutomaticMaskGenerator}(I)$ 
7:    $\mathcal{M} \leftarrow \{m \in \mathcal{M} \mid \text{area}(m) \geq 2000\}$ 
8:    $\mathcal{F} \leftarrow \emptyset$ 
9:   for all  $m \in \mathcal{M}$  do
10:    if ISCONTINUOUS( $m$ ) then
11:       $\mathcal{F} \leftarrow \mathcal{F} \cup \{m\}$ 
12:    end if
13:  end for
14:   $\mathcal{R} \leftarrow \text{SortByScore}(\mathcal{F})$ 
15:  return first  $k$  masks in  $\mathcal{R}$ 
16: end function

17: function ISCONTINUOUS( $mask$ )
18:    $r \leftarrow \text{CHECKROWCONTINUITY}(mask)$ 
19:    $c \leftarrow \text{CHECKCOLUMNCONTINUITY}(mask)$ 
20:   if  $r = \text{False}$  and  $c = \text{False}$  then
21:     return False
22:   else
23:     return True
24:   end if
25: end function

```

▷ $r, c \in \{\text{True}, \text{False}\}$
 ▷ Both directions broken → discard mask
 ▷ At least one direction is continuous → keep

Type	Method	mAP _i	mAP _v	mAP _t	mAP _{ivt}
Conventional	RDV (Nwoye et al., 2022)	34.24	10.77	7.16	2.80
	RiT (Sharma et al., 2023)	27.29	13.08	13.42	3.25
VLM	HecVL (Yuan et al., 2024a)	16.43	14.37	14.08	6.61
	PeskaVLP (Yuan et al., 2024b)	17.62	13.62	14.03	6.88
	SurgVLP (Yuan et al., 2025)	15.29	<u>15.03</u>	<u>15.01</u>	5.52
Ours (Qwen3-VL-8B-Instruct&Thinking)		<u>31.55</u>	26.72	24.11	9.56

Table 3: Performance comparison on the ProstaTD(Chen et al., 2025) dataset with the compositional zero-shot setting.

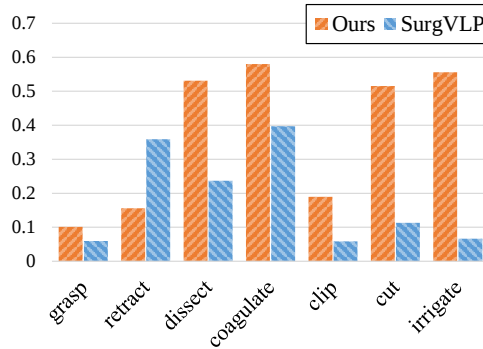


Figure 2: Verb improvements in mAP_v of our framework compared to the SurgVLP (the second-best method).

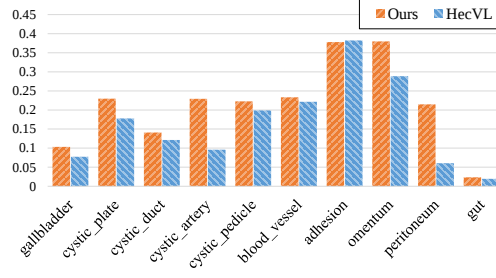


Figure 3: Target improvements in mAP_t of our framework compared to the HecVL (the second-best method).

Glance
User Prompt: This is an intraoperative image captured during laparoscopic surgery. Please carefully observe and provide a detailed paragraph describing the current surgical scene, including the anatomical structures, surgical instruments, and ongoing procedures visible in the image.
Gaze
System Prompt: You are a highly experienced laparoscopic surgeon. Assistant Prompt: [think from glance] + [answer from glance] User Prompt: You are provided with a video clip composed of multiple consecutive frames from the same laparoscopic surgery as the previous image. As the operating surgeon, describe in one paragraph what surgical actions you are currently performing, referring to both the ongoing procedure and any anatomical or procedural context you can observe.
Think
System Prompt: You are a professional laparoscopic surgeon and an expert in surgical scene annotation. Your task is to predict surgical actions based on intraoperative images, strictly following a structured format. Assistant Prompt: [think from glance] + [answer from glance] [answer from gaze] User Prompt: This is an intraoperative image from the same laparoscopic surgery as before. Based on this image, predict the set of actions currently being performed. **Important Instructions**: Each action must be represented as a structured triplet with the following fields: - instrument: must be chosen from the following list: ['grasper', 'bipolar', 'hook', 'scissors', 'clipper', 'irrigator']. - verb: must be chosen from the following list: ['grasp', 'retract', 'dissect', 'coagulate', 'clip', 'cut', 'aspirate', 'irrigate', 'pack', 'null_verb']. - target: must be chosen from the following list: ['gallbladder', 'cystic_plate', 'cystic_duct', 'cystic_artery', 'cystic_pedicle', 'blood_vessel', 'fluid', 'abdominal_wall_cavity', 'liver', 'adhesion', 'omentum', 'peritoneum', 'gut', 'specimen_bag', 'null_target']. - confidence: a float number between 0 and 1 indicating your confidence in the prediction. Only select values from the given lists. Do not invent or hallucinate new labels. If you are uncertain, select the most plausible option based on visual evidence. **Output Format**: Output only a JSON-formatted list of triplets. Do not include any explanations, commentary, or text.

Figure 4: Text prompts used in the Chain-of-Surgery (CoS).

triplet category for validation		triplet category for training	
id	name	id	name
0	<grasper, dissect, cystic-plate>	52	<hook, coagulate, liver>
2	<grasper, dissect, omentum>	55	<hook, cut, peritoneum>
3	<grasper, grasp, cystic.artery>	57	<hook, dissect, cystic.artery>
5	<grasper, grasp, cystic-pedicle>	58	<hook, dissect, cystic.duct>
8	<grasper, grasp, gut>	59	<hook, dissect, cystic-plate>
15	<grasper, retract, cystic-pedicle>	60	<hook, dissect, gallbladder>
24	<bipolar, coagulate, cystic.artery>	61	<hook, dissect, omentum>
25	<bipolar, coagulate, cystic.duct>	62	<hook, dissect, peritoneum>
26	<bipolar, coagulate, cystic-pedicle>	63	<hook, retract, gallbladder>
31	<bipolar, coagulate, peritoneum>	64	<hook, retract, liver>
32	<bipolar, dissect, adhesion>	66	<scissors, cut, adhesion>
35	<bipolar, dissect, cystic-plate>	68	<scissors, cut, cystic.artery>
38	<bipolar, grasp, cystic-plate>	69	<scissors, cut, cystic.duct>
41	<bipolar, retract, cystic.artery>	71	<scissors, cut, liver>
42	<bipolar, retract, cystic-pedicle>	76	<scissors, dissect, omentum>
43	<bipolar, retract, gallbladder>	78	<clipper, clip, cystic.artery>
45	<bipolar, retract, omentum>	79	<clipper, clip, cystic.duct>
46	<hook, coagulate, blood.vessel>	82	<irrigator, aspirate, fluid>
47	<hook, coagulate, cystic.artery>	84	<irrigator, dissect, cystic-pedicle>
48	<hook, coagulate, cystic.duct>	87	<irrigator, dissect, omentum>
49	<hook, coagulate, cystic-pedicle>	88	<irrigator, irrigate, abdominal.wall.cavity>
50	<hook, coagulate, cystic-plate>	90	<irrigator, irrigate, liver>
53	<hook, coagulate, omentum>	92	<irrigator, retract, liver>
54	<hook, cut, blood.vessel>	93	<irrigator, retract, omentum>
56	<hook, dissect, blood.vessel>	94	<grasper, null.verb, null.target>
65	<scissors, coagulate, omentum>	95	<bipolar, null.verb, null.target>
67	<scissors, cut, blood.vessel>	96	<hook, null.verb, null.target>
70	<scissors, cut, cystic-plate>	97	<scissors, null.verb, null.target>
72	<scissors, cut, omentum>	98	<clipper, null.verb, null.target>
73	<scissors, cut, peritoneum>	99	<irrigator, null.verb, null.target>
74	<scissors, dissect, cystic-plate>		
75	<scissors, dissect, gallbladder>		
77	<clipper, clip, blood.vessel>		
80	<clipper, clip, cystic-pedicle>		
81	<clipper, clip, cystic-plate>		
83	<irrigator, dissect, cystic.artery>		
85	<irrigator, dissect, cystic.plate>		
86	<irrigator, dissect, gallbladder>		
89	<irrigator, irrigate, cystic-pedicle>		
91	<irrigator, retract, gallbladder>		

Table 4: Triplet categories splitting list with $\gamma = 80$.